Project Proposal

Rui Shi

May 7, 2019

1 Dataset description and literature review

This dataset comprises of information of motor vehicle collisions in New York City from 2012 to current provided by the New York Police Department. The data would be updated every month. Each row represents a collision in New York city and decribed by time, date, borough, street name, numbers of persons injured and killed, contributing factors and involved vehicle type. New York Police Department collects collision information for each accident and update the data monthly.

The data is open to public for analysing the traffic safety in NYC and exploring related influence factors. The data has been used to explore accident trends and relations between number of collisions and different factors. Dataset link: new-york-city-motor-vehicle-collisions

Since this data contains collision information from 2014 to current. The number of accidents over the years and their breakdown based on boroughs were conducted for making comparison and exploring the trend. The major contributing factors for accidents and accidents related with person injured/killed have been analysed. By filtering the contributing factors based on alcohol involvement and analysing with time factor, the influence of alcohol has been explored. Also, the influence of time factors has been analysed using this data.

2 Questions will be addressed and analysis discussion

For people who live in New York for a long time, it is not surprise that they know several accident black spots and pay more attention while driving near these locations. But for people who rarely drive or just come to this city, it is dangerous to not aware of these facts. The questions that is expected to address:

1. Was the situation different for boroughs in NYC?

2. Was the situation different for different days in a week, hours in a day, month in a year?

3. Was the situation different for different car types?

4. Were there any other factors can influence the situation? (holidays, population...)

By addressing these problems and providing a visualization of the results, people can better understand to what extent these factors can influence the situation and make better choices to avoid accidents. The data contains the date, time and location of each accident, so that problems related with time factors and location can be addressed after properly processing the information. To analysis the trend related with holidays and population, data-sets contain related information need to be merged. The data also includes involved vehicle type information. By using this information and merging data-set contains total number of registered vehicles of every vehicle type, the problem related with car type can be addressed. The influence of time factors, location and vehicle type would be analysed. The relation between time factors and number of collisions has been analysed for the past several years in previous works. The similar analysis would be performed to see if the result would be the same in 2018. However, the influence of holidays has not been analysed. In this project, the trend of the number of collisions in the time period of 3 days before to 3 days after holidays would be explored by merging with another data-set. The breakdown of the number of collisions based on boroughs also has been conducted. This project would continue to explore the relation between population and the number of collisions happened in boroughs. Moreover, the influence of vehicle type would be explored. The data includes the number of register vehicles for each vehicle type would be merged to see the influence clearer. At the end, a prediction of number of person injured/killed would be conducted based on the location, time and other related information.

3 Initial Data Issues

1) This dataset has tons of missing data and empty units. For location identifiers, some offer both borough and GPS location, some leave borough or GPS location empty, some have none of them. For vehicle type identifiers, some collisions involved more than two vehicles, but most only involved two, which leads lots of blank in columns show vehicle information of number 3, 4 and 5.

2) Character data is in different form: street names and vehicle types are mixes of uppercase and lowercase.

3) There are duplicate columns.